СИНТЕЗАТОР РЕЧИ ПО ТЕКСТУ КАК КОМПЬЮТЕРНОЕ СРЕДСТВО “КЛОНИРОВАНИЯ“ ПЕРСОНАЛЬНОГО ГОЛОСА

Лобанов Б.М., Карневская Е.Б., Левковская Т.В.

Институт технической кибернетики НАН Беларуси

Введение.

Основные принципы синтеза русской речи по тексту уже достаточно давно были изложены в ряде работ автора [1]. На их основе разработаны образцы синтезаторов речи серии “Фонемофон”, обеспечивающих качество и разборчивость синтезированной речи вполне пригодное для широкого спектра практических приложений. Однако, не смотря на все усилия, синтезированная речь оставалась ещё далёкой по качеству от натуральной и обладала узнаваемым машинным акцентом. Причиной этому были не столько уровень наших знаний о процессах речеобразования и о фонетике, сколько нехватка вычислительных ресурсов компьютеров того времени. Сейчас мы можем не ограничивать себя ни объёмом оперативной и дисковой памяти, ни требуемым объёмом вычислений и приступить к созданию системы синтеза русской речи с максимально возможным приближением по звучанию к голосу и речи конкретного диктора.

Такая постановка задачи хотя и отдалённо, но напоминает широко известную биологическую проблему клонирования живого существа, когда на основе сравнительно малого объёма генетической информации делается попытка воспроизвести копию живого существа естественно-биологическим путём. В данном случае, в отличие от классической задачи клонирования, делается попытка получения близкой копии не всего существа в целом (в данном случае человека), а только некоторой одной из его функций: чтение произвольного орфографического текста с максимально возможным сохранением персональных акустических особенностей голоса, фонетических характеристик, акцента и просодической индивидуальности речи (мелодика, ритмика, динамика). Кроме того, очевидно, что это всего лишь компьютерный, а не биологический клон, однако обладатели такого “клона” всё же могут быть уверены, что хотя бы частица их сущности - их голос и манера чтения - останутся нетленными.

Общая структура синтезатора (рис. 1).

Входной орфографический текст подвергается ряду последовательных обработок с помощью специальных процессоров. Текстовый процессор предназначен для преобразования входного орфографического текста в размеченный фонемный текст. Под разметкой понимается разбиение текста на отдельные элементы в следующей иерархии: фонетический период, фраза, синтагма. Кроме того, процессор осуществляет: расстановку словесных ударений и интонационную маркировку синтагм. Размеченный фонемный текст поступает на вход 2-х процессоров: просодического и фонетического. В результате работы просодического процессора фонемный текст делится на акцентные группы (АГ). Далее осуществляется разметка АГ на элементы акцентных групп (ЭАГ): интонационное предъядро, ядро и заядро. И наконец, последняя функция просодического процессора - это установка значений амплитуды (А), длительности фонем (Т) и частоты основного тона (F0) для каждого ЭАГ. Задача фонетического процессора заключается в генерации позиционных и комбинаторных аллофонов. Акустический процессор на основе информации о том, какие аллофоны необходимо синтезировать, а также какие просодические характеристики должны быть приписаны каждому аллофону, генерирует речевой сигнал путем компиляции отрезков естественных звуковых волн соответствующих аллофонов.

Рис1. Структурная схема синтезатора речи

2. Фонемно-аллофонное преобразование.

Фонемно-аллофонное преобразование обеспечивает генерацию следующих позиционных аллофонов гласных*: ударный (0), первый предударный (1), не первый предударный (2), заударный (3)*. С учётом левого контекста генерируются следующие комбинаторные аллофоны гласных: после *синтагматической паузы (0), после переднеязычных (1), губных (2) и заднеязычных (3) твёрдых, после /Л/ (4), после /Р/ (5), большинства мягких (6), после /Р'/ (7), после /М'/ (8), после /Н'/ (9), после гласных (У), (О), (А), (Э), (Ы), (И)*. Всего - 16 левых контекстов. С учётом правого контекста генерируются следующие комбинаторные аллофоны гласных: перед синтагматической паузой (0), перед переднеязычными и заднеязычными (1) и перед губными (2) твёрдыми, перед мягкими (4). Итого, в общем случае, обеспечивается генерация Nv = 4*16*5*6(гласных) = 1920 гласных аллофонов. Их число, реально используемое в синтезаторе с учётом известных закономерностей, – менее 1000.

Аллофоны согласных генерируются с учётом левого и правого контекста. Левый контекст: *после паузы (0), после глухих (1) и звонких (2) согласных, после гласных (3)*. Правый контекст: *перед паузой (0), перед глухими (1) и звонкими (2) согласными, перед безударными (3) и ударными (4) гласными*. Итого, в общем случае, обеспечивается генерация Nc = 4*5*36(согласных) = 720 согласных аллофонов. Их число, реально используемое в синтезаторе с учётом известныхограничений, – менее 500.

Подробная схема генерации аллофонов гласных и согласных фонем представлена на рис. 2.

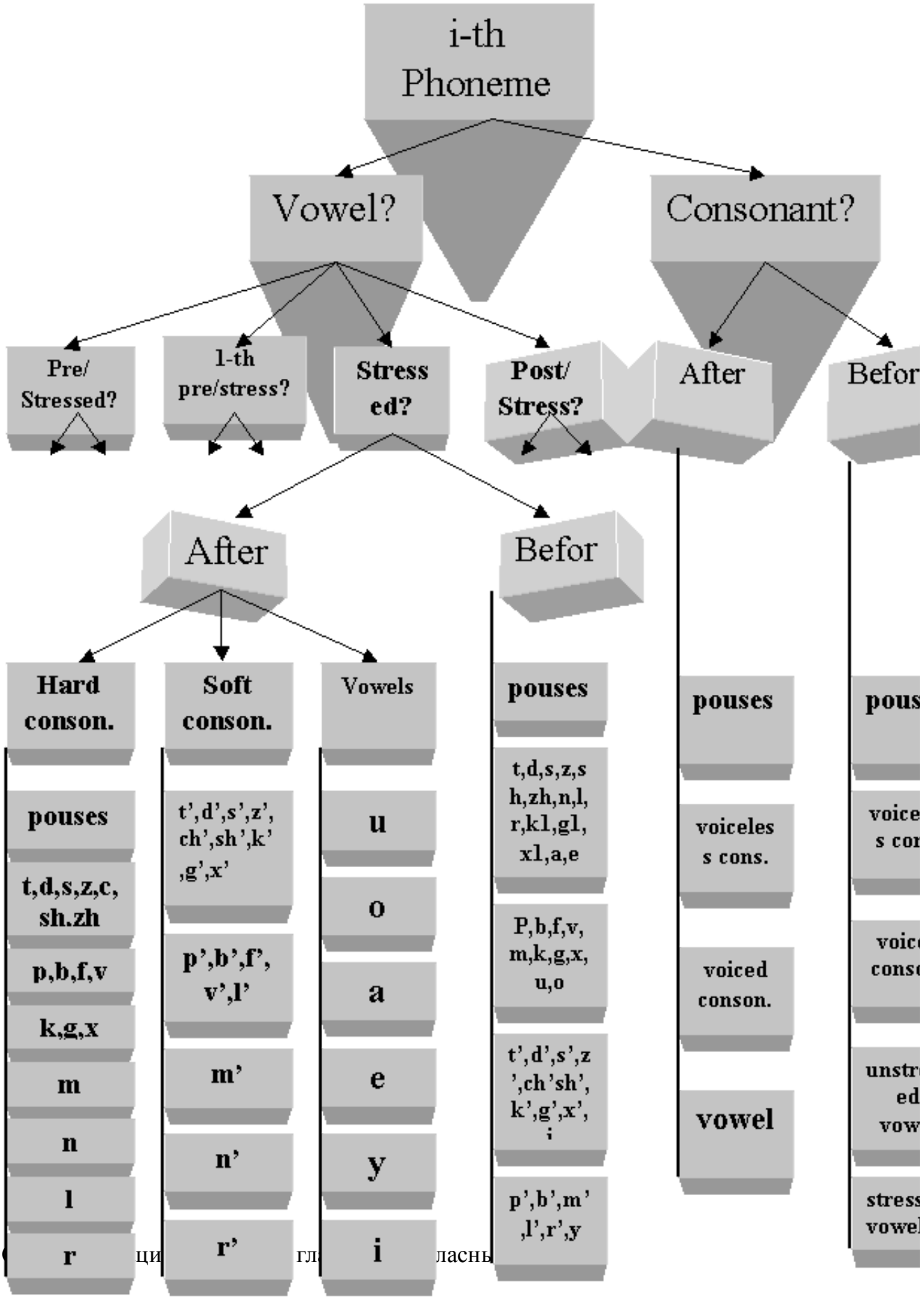

Рис. 2. ци гл ласны

3. Стратегия персонализации синтезированной речи.

Основой успешного решения задачи персонализации звучания синтезированной речи является корректное выполнение следующих трёх требований:

1. Максимально полное использование при синтезе речи акустических, фонетических и просодических средств индивидуальности голоса и речи имитируемого диктора;

2. Минимально возможные искажения элементов компиляции в процессах их записи, воспроизведения и просодической модификации.
3. Отсутствие дополнительных преобразований записанных сигналов, таких как преобразование Фурье или PSOLA.

Персональные акустические особенности голоса диктора сохранены благодаря высококачественной цифровой записи элементов компиляции (аллофонов). Персональные фонетические и просодические характеристики речи сохранены благодаря максимально полному копированию их проявления в реальной речи диктора. Использование отрезков натурального речевого сигнала (звуковых волн), точная разметка концов аллофона и его питчей (периодов), “щадящая” модификация ЧОТ обеспечили минимальные искажения элементов компиляции при их воспроизведении.

4. База данных звуковых волн аллофонов

База данных звуковых волн аллофонов хранится в виде сигналов в Wav-формате с частотой дискретизации 16 кГц и разрядностью 16 бит. Общее количество Wav-файлов – 1100, объём занимаемой памяти – 4,2 МВ.

Каждый Wav-файл сопроваждается заголовком, в котором указаны:

- имя аллофона (три символа, например A132),
- число отсчётов сигнала – N,
- число питчей (периодов) – P,
- позиция каждого питча в номерах отсчётов сигнала – p1,p2,…pP,
- позиция срединного питча аллофона – ps,
- амплитуда аллофона – A.

На рис. 3 представлен сигнал, соответствующий аллофону A132, на котором указаны положения каждого питча.

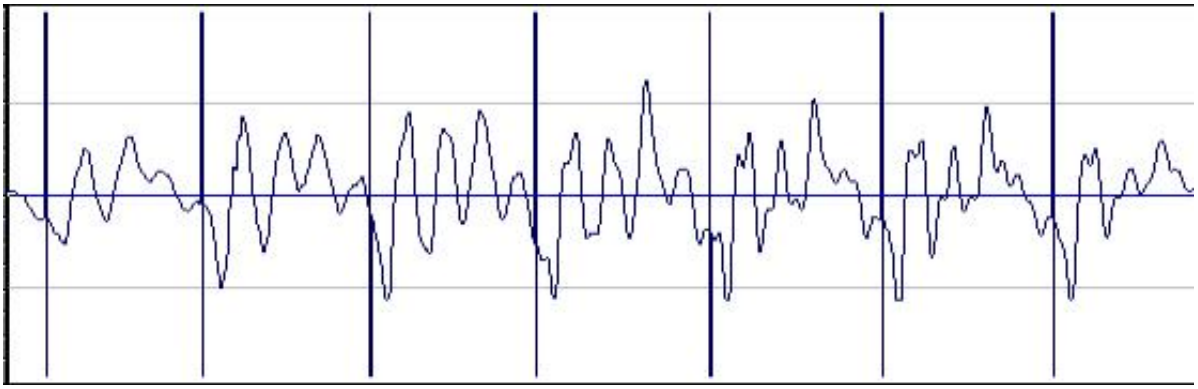

Рис3. Сигнал, размеченный на питчи

5. Модификация частоты основного тона - F0

Модификация частоты основного тона осуществляется путём изменения длительности текущего периода основного тона звуковых волн аллофонов: удлиннения - при уменьшении F0, или их укорочения - при увеличении F0 (см. Рис. 4, 5).

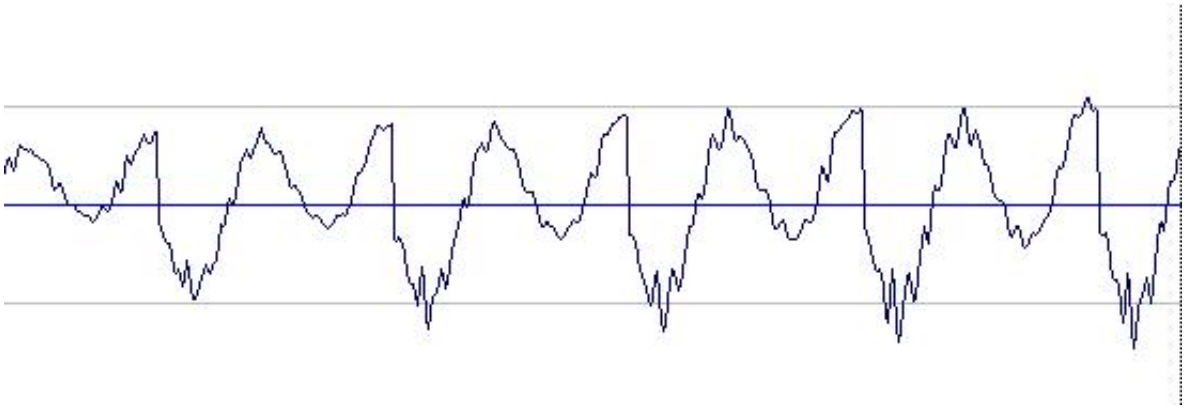

Рис.4. Пример укорочения периодов

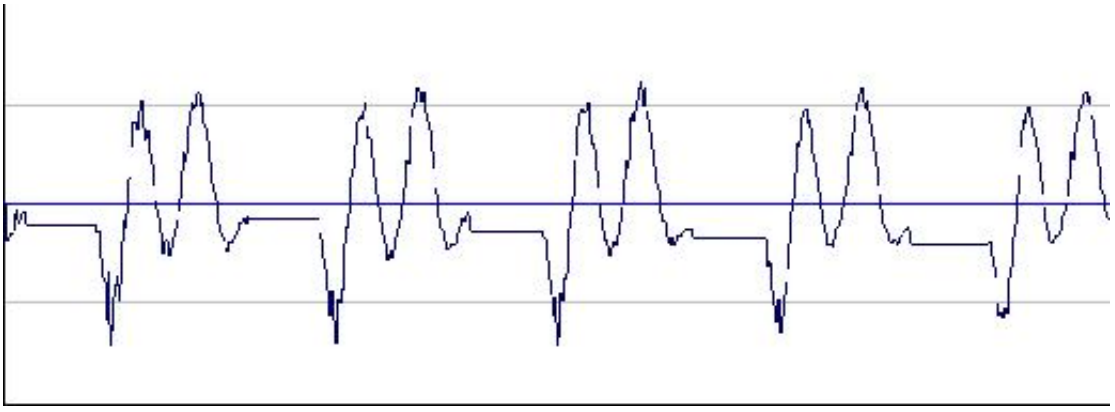

Рис. 5. Пример удлиннения периодов

При этом, как видно из рис. 4 и 5, могут возникать весьма чувствительные для слуха искажения речевого сигнала. Чтобы предотвратить этот эффект осуществляется локальное сглаживание левой (i-й) и правой (j-й) стыкуемых волн по следующему алгоритму.

А) При укорочении периода.

1. От последнего (нулевого) отсчёта левой (i-й) стыкуемой волны отсчитываем 3-й отсчёт, для которого расчитываем новое среднее значение Si3m из значений i-й и j-й волн по формуле:

$$Si3m = 1/9 * (Si7+...+Si3...+ Si0+Sj0)$$

2. Далее процесс повторяется по следующей рекуррентной схеме вплоть до получения последнего нового значения для 0-го отсчёта i-й волны:

$$Si2m = 1/9 * (Si6+...+Si2...+ Sj0+Sj1)$$
$$Si1m = 1/9 * (Si5+...+Si1...+ Sj1+Sj2)$$
$$Si0m = 1/9 * (Si4+...+Si0...+ Sj2+Sj3)$$

3. Затем рассчитываются новые значения j-й волны:

$$Sj0m = 1/9 * (Si3+... Sj0m... + Sj3+Sj4)$$
$$Sj1m = 1/9 * (Si2+... Sj1m... + Sj4+Sj5)$$
$$Sj2m = 1/9 * (Si1+... Sj2m... + Sj5+Sj6)$$

4. Процесс заканчивается после получения нового значения для 4-го отсчёта j-й волны:

$$Sj3m = 1/9 * (Si0 + Sj0 + \ldots Sj3m \ldots + Sj6 + Sj7)$$

Б) При удлиннении периода алгоритм сглаживания аналогичен. Сохраняются также условия начала и конца процесса сглаживания. Добавляется лишь дополнительный k-й участок, повторяющий значение последнего (0-го) отсчёта i- волны, который вначале играет роль j-го участка, а затем i-го. Иначе говоря, единообразный алгоритм обеспечивает локальное сглаживание, начиная с 3-го отсчёта i–й волны и кончая 3–м отсчётом j–й волны независимо от присутствия или отсутствия дополнительного участка.

3. Программная реализация синтезатора.

Синтезатор речи реализован в среде визуального программирования Microsoft Visual C++ 6.0 для операционных систем Windows 98/NT. Минимальные технические требования к компьютеру: 7Mb свободного места на жестком диске, 166MHz процессор, 32Mb оперативной памяти.

Система состоит из набора компонентов, соответствующих процессорам, базе данных расстановки ударений и базе данных аллофонов. База данных с ударениями находится в библиотеке динамической компоновки, что позволяет быстро загружать и выгружать её при инициализации и завершении работы системы. Поиск ударения в словаре реализован алгоритмом бинарного поиска. Для оптимизации поиска нужного аллофона, аллофоная база загружается в память при инициализации системы. Текст анализируется синтагмами: после обнаружения синтагмы, в ней автоматически расставляются ударения, далее осуществляется букво-фонемное преобразование, а затем фонемно-аллофоное. Параллельно идентифицируется следующая синтагма, которая подвергается аналогичнойобработке. По типу синтагмы определяется её просодические характеристики. Как только синтагма сформирована, она проговаривается и одновременно обрабатывается следующая за ней синтагма. Таким образом практически не заметна задержка между формированием синтагм.

Заключение. Первые результаты по “клонированию речи” касались только синтеза мужского голоса [2]. Работа синтезатора речи с мужским и женским персонализированными голосами будет продемонстрирована во время доклада на конференции. Участникам конференции мы предоставляем возможность самостоятельно судить насколько успешно удалось синтезировать персональные особенности голоса и речи авторов этого доклада.

Литература

1. Лобанов Б.М. Ретроспективный обзор исследований и разработок Лаборатории распознавания и синтеза речи. Сб. “Автоматическое распознавание и синтез речи”, ИТК НАН Беларуси, Минск, 2000.-С.6-23.
2. Киселёв В.В, Левковская Т.В.,Лобанов Б.М.,Хейдоров И.Э. Синтезатор персонализированной речи по тексту*”ЛобаноФон-2000”*. Тр. Международной конференции “100 лет экспериментальной фонетике в России”, Ст.-Петербург, 2001.-С.101-104.